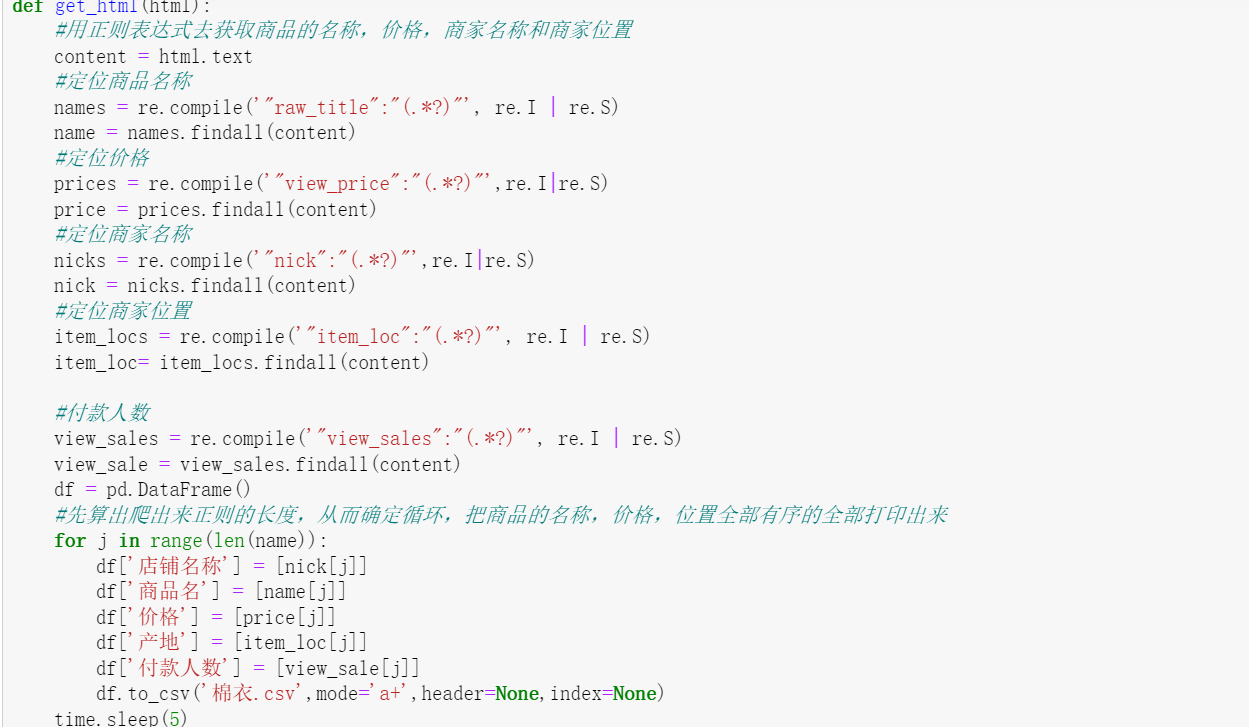
首先第一步我们是先去爬取数据，在这里我们是选用中国最大的电商网站淘宝网

但是因为淘宝网反爬机制很强，所以我们就意思意思一下爬10页数据就好了



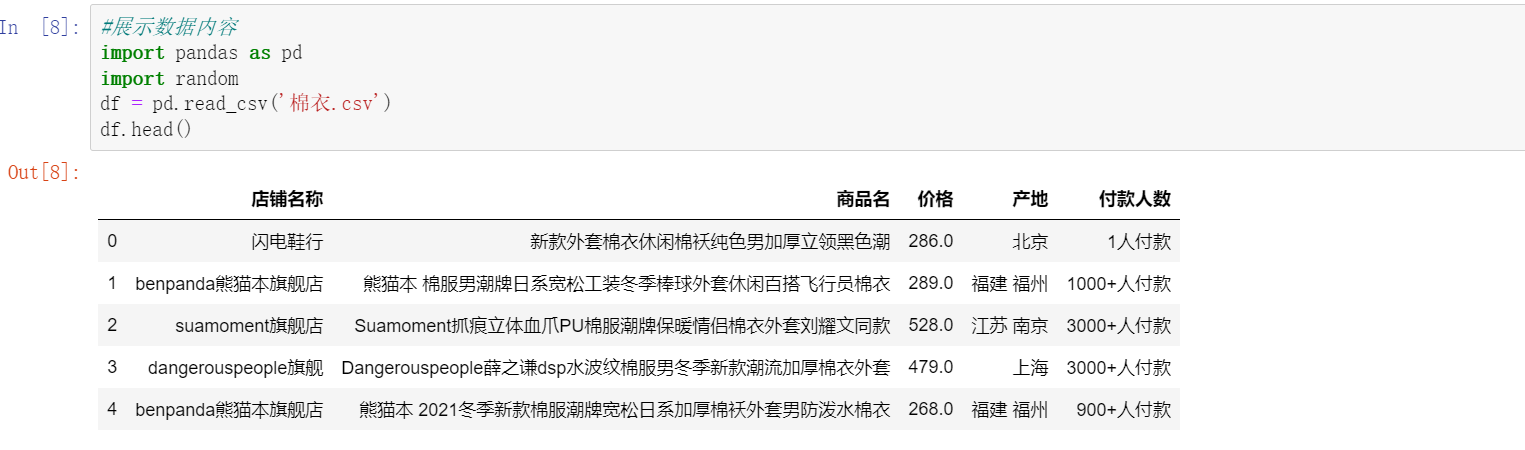
在这里我们用正则表达式去定位数据的内容

然后爬完一页停留5秒再继续爬，防止被网站检测出是爬虫程序

然后爬完数据如下



接着我们再去看我们的CSV文件



在这里我们展示了前5条数据

该有的内容都在里面

然后我们开始一步步去分析

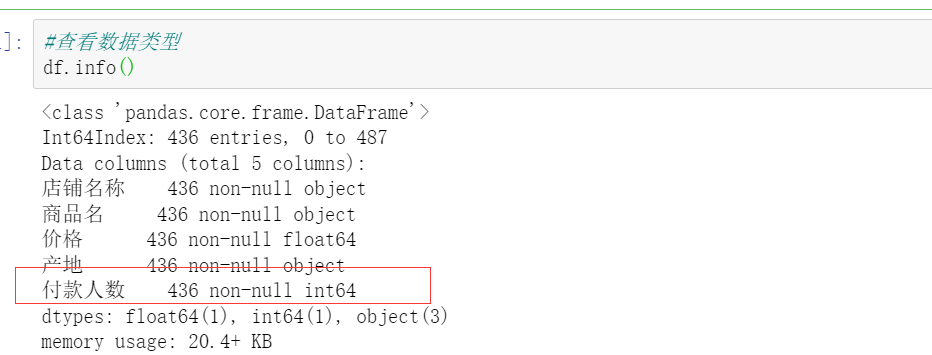
首先我们先把付款人数里面多余的字符去掉里面xxx付款后面的付款就不要了，只保留数据便于我们去分析内容



先删掉重复行

然后用一个函数把这些字符去掉

然后再去查看我们的数据



这时付款人数已经从object变成int说明是全部清洗干净了

然后我们再对商品的名称进行数据清洗



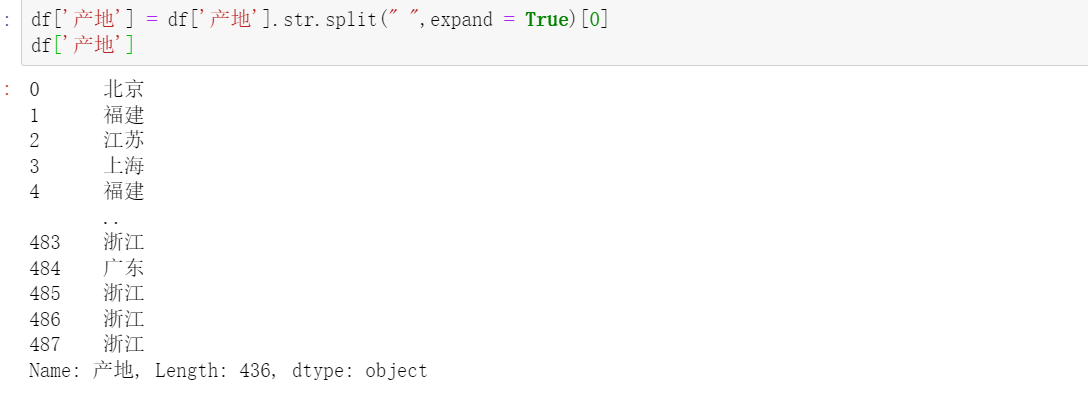
清洗一些重复数据或者表情包这种影响分析的数据

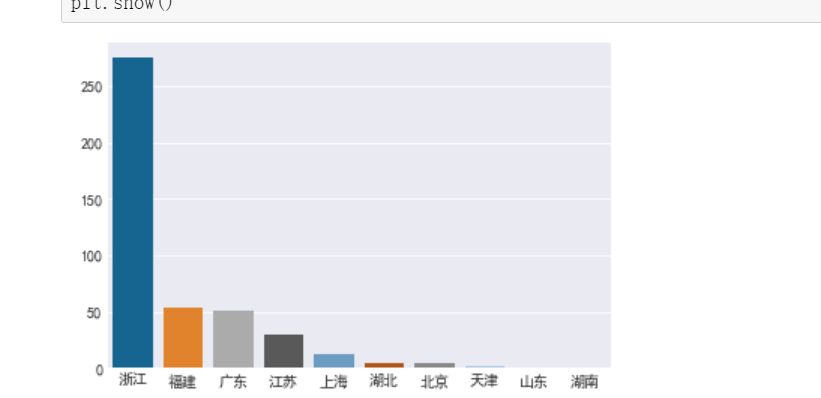
接着我们再用我们的jieba库和词云库对商品名称进行一个分词和可视化的展示



从这里不难看出棉衣的热词是 棉衣 棉服 外套 宽松 男生 冬季 潮牌 加厚 棉袄 这些关键词为主

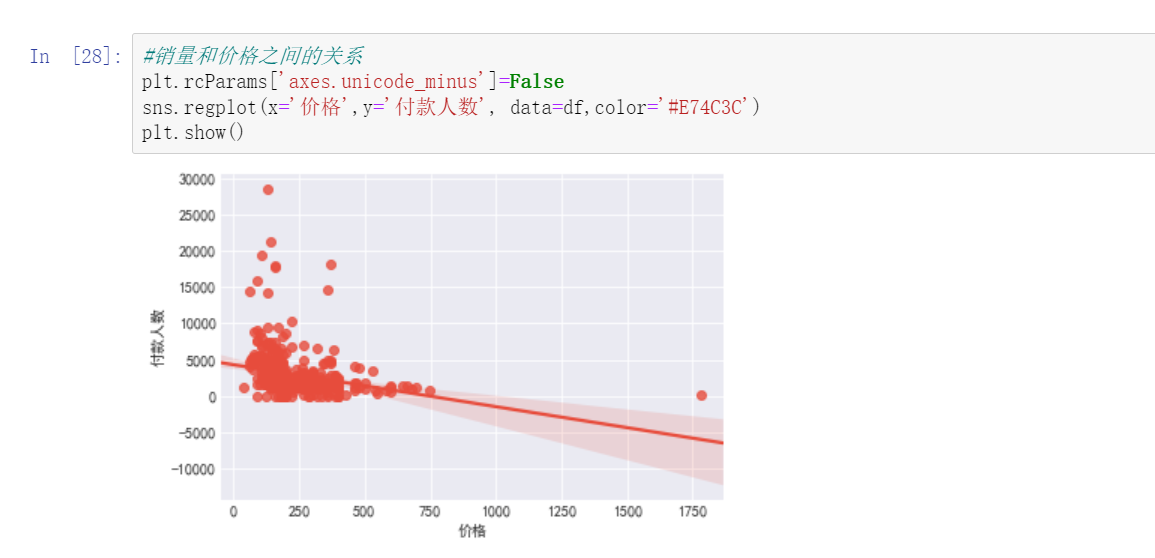
然后我们再去看他们的产地集中在哪一块





从柱状图我们可以清晰的了解到，主要产地还是以沿海地区为主，然后浙江是遥遥领先其他产地，这也和浙江是电商之乡有关

然后我们再去做一个线性回归分析

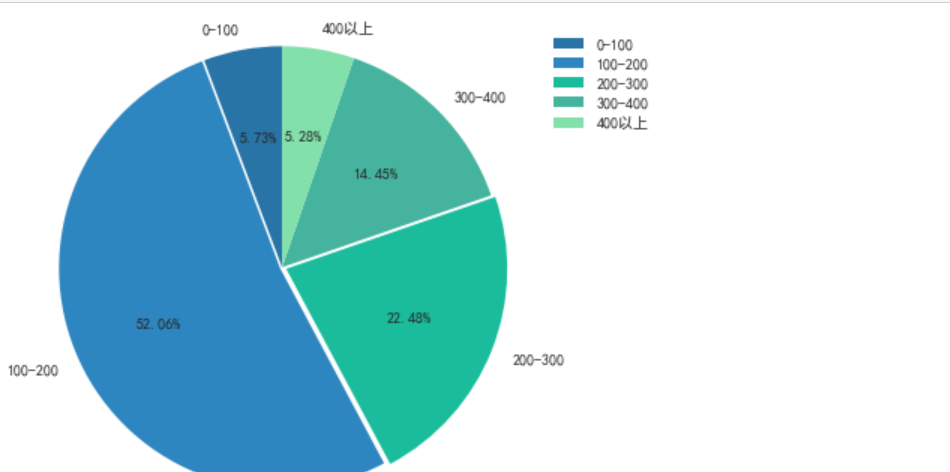


果然不出所料，价格和销量成反比，成一个负相关，这也符合我们的日常，东西越贵，购买的人越少，销量自然越低，价格越便宜销量自然越高

最后根据上面的内容我们做一个价格规划



看看哪一块的价格占比最高



在0-300这个区间占比最高，说明人们还是追求经济实惠为主